Proper Evaluation of Neural Network and Learning Systems based Prediction Intervals

Pierre Pinson, Senior Member, IEEE, Julija Tastu

Abstract—Neural Network and Learning Systems approaches are increasingly used in probabilistic prediction. Forecast evaluation then comprises a complex task for which a number of scores have been proposed, aiming to summarize the assessment of their overall quality with a single number. Such scores ought to be proper though, that is, to effectively reward interval forecasts of higher quality. If not, the ranking of score values does not allow concluding on the actual superiority of a given approach over others, since one may always have the possibility to hedge in order to obtain the best score value. Recently, the Coverage Width-based Criterion (CWC) was proposed and used for an evaluation of the state of the art. The CWC score is shown to be improper based on theoretical considerations, while the consequences are explored.

Index Terms—prediction, interval forecasts, verification, skill, proper score

I. INTRODUCTION

In different areas of forecasting, substantial developments are on proposing and applying probabilistic approaches, as for instance in economics and finance [1], meteorology [2], as well as in various aspects of power systems management e.g. electric load [3], wholesale market prices [4] and renewable energy production [5]. Among the alternative methodologies for probabilistic prediction, approaches based on neural networks and more generally learning systems have gained increased interest over the last two decades, from the early work of [6] to the recent review in [7].

Prediction intervals (also referred to as interval forecasts) are some of the probabilistic forecasts for continuous variables that attracted the most attention since providing visual and easily interpretable information about forecast uncertainty. Rigorous methodologies are required for their evaluation and for the comparison of rival approaches. For examples of benchmark exercises, see [8], [9] among others. Evaluation and comparison are to be based on proven scores and diagnostic tools. Such scores ought to be proper [10], [11]: propriety is the basic property of a score to ensure that perfect forecasts should be given the best score value, say, the lowest one if the score is negatively oriented. It appears that such a crucial aspect is not always respected.

Emphasis is placed here on the recent proposal of the Coverage Width-based Criterion (CWC) [9] as a score for the comparison of Neural Network and Learning Systems (NNLS) approaches to issuing prediction intervals. The necessary background on proper scores for prediction intervals is first introduced in Section II. Subsequently, our original contribution consists in (i) showing in Section III that the CWC score is not proper, and (ii) underlining in Section IV the consequences of employing this improper score, based on the example of a simple hedging strategy permitting to always obtain score values better than those of rival approaches. Finally in Section V, we conclude on the fact that, owing to the lack of property, the ranking of CWC score values does not allow concluding on the actual superiority of a given approach over another, shedding doubts on the evaluation of the state of the art performed in [9].

II. PROPER SCORES FOR PREDICTION INTERVALS

Probabilistic forecast verification frameworks were proposed over the last 30 years, their main principles being underlined in [11], [13], [14]. They involve the evaluation of calibration (correspondence of nominal and empirical probabilities), as well as sharpness (concentration of probability—the tighter the better) using a set of diagnostic tools and scores.

Consider a stochastic process \( \{ G_t \} \) for which interval forecasts are to be generated, say, for a lead time \( k \): \( I_{\ell+k|t}^{(\beta)} \) are central prediction intervals with nominal coverage rate \((1-\beta)\), issued at time \( t \) for lead time \( t+k \),

\[
I_{\ell+k|t}^{(\beta)} = \left[ q_{\ell+k|t}^{(\alpha)}, q_{\ell+k|t}^{(\beta)} \right],
\]

where \( q_{\ell+k|t}^{(\alpha)} \) and \( q_{\ell+k|t}^{(\beta)} \) are quantile forecasts whose nominal levels \( \alpha \) and \( \beta \) are quantile forecasts whose nominal levels \( \alpha \) and \( \beta \), such that \( \alpha = 1 - \beta = \beta/2 \).

The joint assessment of calibration and sharpness ideally relies on a unique criterion, a score \( Sc \), that assigns a single value \( Sc(I_{\ell+k|t}^{(\beta)}; y_{t+k}) \) to each forecast-observation pair, then to be averaged over an evaluation set, \( t = 1, \ldots, T \). If knowing \( \{ G_t \} \), perfect interval forecasts \( I_{\ell+k|t}^{(\beta)_s} \) would be

\[
I_{\ell+k|t}^{(\beta)_s} = \left[ q_{\ell+k|t}^{(\alpha)}, q_{\ell+k|t}^{(\beta)} \right],
\]

with \( q_{\ell+k|t}^{(\alpha)} \) and \( q_{\ell+k|t}^{(\beta)} \) the actual quantiles of the process at time \( t+k \). Throughout the paper, the "*"-symbol will be associated to perfect forecasts and their score value.

Following [10], a score \( Sc \) for prediction intervals is said to be proper if for any prediction interval \( I_{\ell+k|t}^{(\beta)} \) and corresponding observation \( y_{t+k} \),

\[
Sc(I_{\ell+k|t}^{(\beta)_s}; y_{t+k}; \theta) \leq Sc(I_{\ell+k|t}^{(\beta)}, y_{t+k}; \theta), \quad \forall t, k, \beta, \theta
\]
i.e., perfect interval forecasts $\tilde{I}_{t+k|t}$ are to be assigned the lowest possible score value. Better, a score for intervals is strictly proper if having a strict inequality in (3).

Employing proper scores ensures consistency in forecast verification and when comparing rival approaches. Perfect interval forecasts are obviously not available when focusing on real-world processes. It is hence impossible to have them as a reference the other approaches should try to get close to. However, the mere idea that already from theoretical considerations perfect forecasts would not be given the optimal score value can only bring discredit on the ranking of rival forecasting methods. Worse, as will be shown and discussed in the following, it discourages competitors to issue their best forecasts, instead turning into an incentive to hedge by playing the score.

Following the pioneering work of [15], it was shown (see e.g. [12]) that a family of proper scores for interval forecasts can be readily obtain from scoring rules for its defining quantiles. For instance, the proper score proposed by [15] is defined as

$$\text{Sc} \left( \tilde{I}_{t+k|t}, y_{t+k} \right) = \left( \hat{q}_{t+k|t}^{(\pi)} - \hat{q}_{t+k|t}^{(\sigma)} \right) + \frac{2}{\beta} \left( \hat{q}_{t+k|t}^{(\alpha)} - y_{t+k} \right) \mathbf{1} \{ y_{t+k} < \hat{q}_{t+k|t}^{(\alpha)} \} + \frac{2}{\beta} \left( y_{t+k} - \hat{q}_{t+k|t}^{(\pi)} \right) \mathbf{1} \{ y_{t+k} \geq \hat{q}_{t+k|t}^{(\pi)} \}.$$  

(4)

$$\hat{\xi}_{t,k} = \mathbf{1} \{ y_{t+k} \in \tilde{I}_{t+k|t}^{(\beta)} \} = \begin{cases} 1, & \text{if } y_{t+k} \in \tilde{I}_{t+k|t}^{(\beta)}, \\ 0, & \text{otherwise} \end{cases}$$  

(5)

It naturally rewards sharp intervals, while penalizing cases where the observation is not covered. It is to be averaged over an evaluation set, $t = 1, \ldots, T$.

### III. THE COVERAGE WIDTH-BASED CRITERION IS NOT A PROPER SCORE

#### A. Definition of the Coverage Width-based Criterion

The calibration assessment of interval forecasts follows a frequentist approach, by comparing their empirical and nominal coverage rates. The empirical coverage rate is derived based on the indicator variable $\xi_{t,k}$, defined for a prediction interval $\tilde{I}_{t+k|t}^{(\beta)}$ and corresponding observation $y_{t+k}$ as

$$\tilde{\xi}_{t,k} = \mathbf{1} \{ y_{t+k} \in \tilde{I}_{t+k|t}^{(\beta)} \} = \begin{cases} 1, & \text{if } y_{t+k} \in \tilde{I}_{t+k|t}^{(\beta)}, \\ 0, & \text{otherwise} \end{cases}$$  

(5)

$$b_k = \frac{1}{T} \sum_{t=1}^{T} \tilde{\xi}_{t,k}.$$  

(6)

The difference $\Delta b_k = (1 - \beta) - b_k$ between nominal and empirical coverage rates is to be seen as a probabilistic bias of the prediction intervals, also referred to as calibration deficit.

Since for probabilistically calibrated intervals, sharpness is a desired property, one may monitor the width of prediction intervals and then derive some summary statistics. The average prediction interval width over an evaluation set of length $T$ is

$$\overline{\delta}_k = \frac{1}{T} \sum_{t=1}^{T} (\hat{q}_{t+k|t}^{(\sigma)} - \hat{q}_{t+k|t}^{(\alpha)}).$$  

(7)

Using these measures for probabilistic calibration and sharpness, the CWC score, for a give lead time $k$ and nominal coverage rate $(1 - \beta)$, was introduced by [9] as

$$\text{CWC} = \tilde{\delta}_k (1 + \mathbf{1} \{ \Delta b_k > 0 \} \exp \left( \eta \Delta b_k \right)), \quad \eta > 0,$$  

(8)

This score penalizes interval forecasts if their empirical coverage rate is lower than the nominal one, while it rewards sharpness otherwise. At first sight it could be seen as similar in essence to the score in (4).

#### B. Why is the CWC score not proper?

If the CWC score were proper, there could not be any set of interval forecasts getting a CWC score value lower than that of perfect interval forecasts. In view of (8), the CWC score value for perfect interval forecasts would simplify to their average width. Let us denote by $\overline{\delta}_k^*$ that optimal score value, with

$$\overline{\delta}_k^* = \frac{1}{T} \sum_{t=1}^{T} \hat{q}_{t+k|t}^{(\sigma)} - \hat{q}_{t+k|t}^{(\alpha)}.$$  

(9)

In the case where one would want to get a better score than $\overline{\delta}_k^*$, the only way is to issue sharper forecasts, since wider intervals can only lead to higher CWC score values anyway. Sharpening the prediction intervals necessarily comes at the expense of calibration, since empirical coverage would get lower than $(1 - \beta)$, yielding a calibration deficit $\Delta b_k > 0$.

Hedging the CWC score in order to obtain a value lower than $\overline{\delta}_k^*$ translates to

$$\tilde{\delta}_k (1 + \exp \left( \eta \Delta b_k \right)) < \overline{\delta}_k^* \quad \eta > 0,$$  

(10)

which, after a little algebra, yields the following inequality on the free parameter $\eta$,

$$0 < \eta < \frac{\ln \left( \overline{\delta}_k^* / \overline{\delta}_k \right)}{\Delta b_k}, \quad \text{with } \frac{\ln \left( \overline{\delta}_k^* / \overline{\delta}_k \right)}{\Delta b_k} > 0.$$  

(11)

The above inequality demonstrates that in principle, there always exists a value of $\eta$ such that some forecasts could be given a better score than that for perfect forecasts, since $\overline{\delta}_k^*/\overline{\delta}_k > 1$ (otherwise one would not have sharper forecasts), while the calibration deficit $\Delta b_k$ is finite, $\Delta b_k \in [0, (1 - \beta)]$.

Based on inequality (11), some might say that the CWC is a conditionally proper score, where one simply has to pick a value for $\eta$ high enough to ensure no imperfect prediction intervals could get a CWC score lower than $\overline{\delta}_k^*$. However, please consider the limit of the right-hand side quantity in inequality (11) as $\overline{\delta}_k$ tends towards infinity,

$$\lim_{\overline{\delta}_k \to \infty} \frac{\ln \left( \overline{\delta}_k^* / \overline{\delta}_k \right)}{\Delta b_k} = +\infty.$$  

(12)

This is since the numerator necessarily tends towards infinity while the denominator stays finite.

As a consequence, whatever the value chosen for $\eta$, any interval forecasts in the form of a probability mass would be
given a CWC score values better than $\bar{\delta}^c_k$. Looking again at the
definition of the CWC score itself, one indeed observes that
\begin{align}
CWC & \geq 0, \quad (13) \\
CWC & = 0, \quad \text{if } \bar{\delta}^c_k = 0. \quad (14)
\end{align}

With this score, no one can outperform prediction intervals
defined as probability masses on a single value, whatever this
value. As an illustrative example, simply predict your favorite
number all the time, and you will win any forecast comparison
or competition where the CWC is used as the lead score.

IV. ILLUSTRATING THE CONSEQUENCES OF NOT USING A
PROPER SCORE

Let us look at this issue of not employing a proper score
in a more practical manner here, by concentrating on the
example of simple hedging strategies in a benchmark exercise
or forecast competition setup where the CWC is used as the
lead score. In a such a case the free parameter $\eta$ would be
fixed. For instance in the case of Ref. [9], $\eta = 50$.

A competitor knowing the behavior of the CWC score
underlined in the above and aiming to win the benchmark
exercise, will not be tempted to issue his best forecasts, but
instead to hedge in order to obtain a better score value. The
forecast intervals of the best performing competitor consist in
a set of prediction intervals $\{I_{t+k|t} \}$, with average width $\bar{\delta}^c_k$
and calibration deficit $\Delta b^c_k$. Their CWC score is
\begin{equation}
\bar{\delta}^c_k \{1 + 1\{\Delta b^c_k > 0\} \exp(\eta \Delta b^c_k)\}.
\end{equation}

Consequently, the hedging strategy is simply to sharpen such
intervals. To keep it simple, consider a linear scaling factor
to be applied to every individual prediction interval. Similarly
to (10) and after a little algebra, a simple condition for the
CWC score value of any interval to be lower than that of the
best performing interval is
\begin{equation}
\delta_k < \nu \bar{\delta}^c_k, \quad (16)
\end{equation}
where
\begin{equation}
\nu = \frac{1 + 1\{\Delta b^c_k > 0\} \exp(\eta \Delta b^c_k)}{1 + 1\{\Delta b^c_k > 0\} \exp(\eta \Delta b^c_k)}.
\end{equation}

From the above, since the numerator is necessarily greater than
1, while $\Delta b^c_k \leq (1 - \beta)$, one has
\begin{equation}
\nu \geq \frac{1}{1 + \exp(\eta(1 - \beta))},
\end{equation}
where this bounding value is independent of the intervals
themselves and of the potential calibration deficit. It is hence
straightforward to find a maximum value for the scaling factor.
As long as one defines prediction intervals such that
\begin{equation}
\delta_k < \frac{1}{1 + \exp(\eta(1 - \beta))} \bar{\delta}^c_k,
\end{equation}
the CWC score of that competitor will be best, without having
made any attempt to issue high-quality prediction intervals.
Note that even if the best performing interval forecasts were
not known and shared among participants, the best strategy is
simply to issue prediction intervals as sharp as possible, even
if leading to an obvious calibration deficit.

V. CONCLUSIONS

NNLS-based approaches have a great role to play in the
development of probabilistic forecasting methodologies, for a
wide range of applications of industrial and societal relevance.
It is of utmost importance, however, for the state of the art
to progress and to be regularly evaluated on a solid theoretical
basis. This translates to the mandatory usage of proper scores,
which can leave no doubt on the meaning of score values (and
(corresponding ranking of rival approaches) in forecast
competition and benchmark exercises. Our aim here was to insist
on this aspect, based on the recent example of the improper
CWC score. Without this basic propriety requirement, it is
unfortunately impossible to validate the evaluation of the state
of the art in interval forecasting performed with this score as a
lead criterion in [9], and to trust any other forecast comparison
that would rely on the CWC. It is therefore suggested that
future work focusing on probabilistic forecasting with any
form of artificial intelligence and machine learning approaches
rely on proper scores only.

REFERENCES

prediction intervals for load forecasting problem,” IEEE Transactions on
interval forecasting of the electricity price,” IEEE Transactions on Power
wind power generation,” IEEE Transactions on Power Systems, vol. 25,
neural networks,” Journal of the American Statistical Association, vol. 92,
A review of applications,” Expert Systems with Applications, vol. 36, no. 1,
p. 2-17, 2009.
methods for neural networks: a practical comparison,” IEEE Transactions on
review of Neural Network-based prediction intervals and new
advances,” IEEE Transactions on Neural Networks, vol. 22, pp. 1341–
1356, 2011.
importance of being proper,” Weather and Forecasting, vol. 22, pp. 382–
calibration and sharpness,” Journal of the Royal Statistical Society B,
and estimation”, Journal of the American Statistical Association, vol. 102,
with applications to financial risk management,” International